EDA Assignment

Risk Analytics in Banking and Financial Institutions

Upgrad and IIIT Bangalore, Data Science Program

SHRAWANI DAS, DS50

Table of Contents

01. Problem Statement

Given data about loan applications from a consumer finance company which specialises in lending various types of loans to urban customers, using EDA to analyse the patterns present in the data and deliver solutions to minimise the risk of losing money while lending to customers.

02 . How Is Data Analytics Used In Finance And Banking Sector?

1. The increasing interest in the use of data analytics in the finance and banking industry is due to the increased changes in technology, changes in people's expectations, and changes in market structure and behavior.

2. The advent of data analytics have helped these sectors to optimize processes and streamline its operations, thus improving efficiency and competitiveness.

3. Examples of how banks and financial institutions use data analytics to manage risk.
    1. Fraud detection
    2. Risk modelling for investment banks
    3. Credit risk analysis
    4. Operational & liquidity risk

03. Data sets used in this assignment:

1. 'application_data.csv' contains all the information of the client at the time of application.
The data is about whether a **client has payment difficulties.**

2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

3. 'columns_description.csv' is data dictionary which describes the **meaning of the variables**.

04. Steps Involved in Exploratory Data Analysis :

1. Data Collection: Here we have loaded 'application_data.csv, previous_application.

2. 2. Data Understanding: By reviewing columns_description.csv and using functions like describe(), info()

3. Data Cleaning
   - Removing the columns which have null values more than 50% and unnecessary rows/ columns.
   - Re-indexing and reformatting our data i.e changing –ve values in dates to +ve ones, filling error data with specified values.

04. Steps Involved in Exploratory Data Analysis:

4. Handling Outliers: Using Boxplot and filling the null values precent in the remaining columns which have <50% with Mean/Median/Mode respectively.

5. Segmentation: Categorizing columns as numerical, categorical, and unique ones to analyze efficiently.

6. Checking Data imbalance in Target variable: Here Target variable is “TARGET” where "1" has a very high number of observations and "0" has a very low number of observations.

7. Analysis
   a) Univariate Analysis and graphical representation
   b) Bivariate Analysis and graphical representation

** The steps from 3 -7 are done separately; First we handled Data Application and then Previous Data Application**

04. Steps Involved in Exploratory Data Analysis:

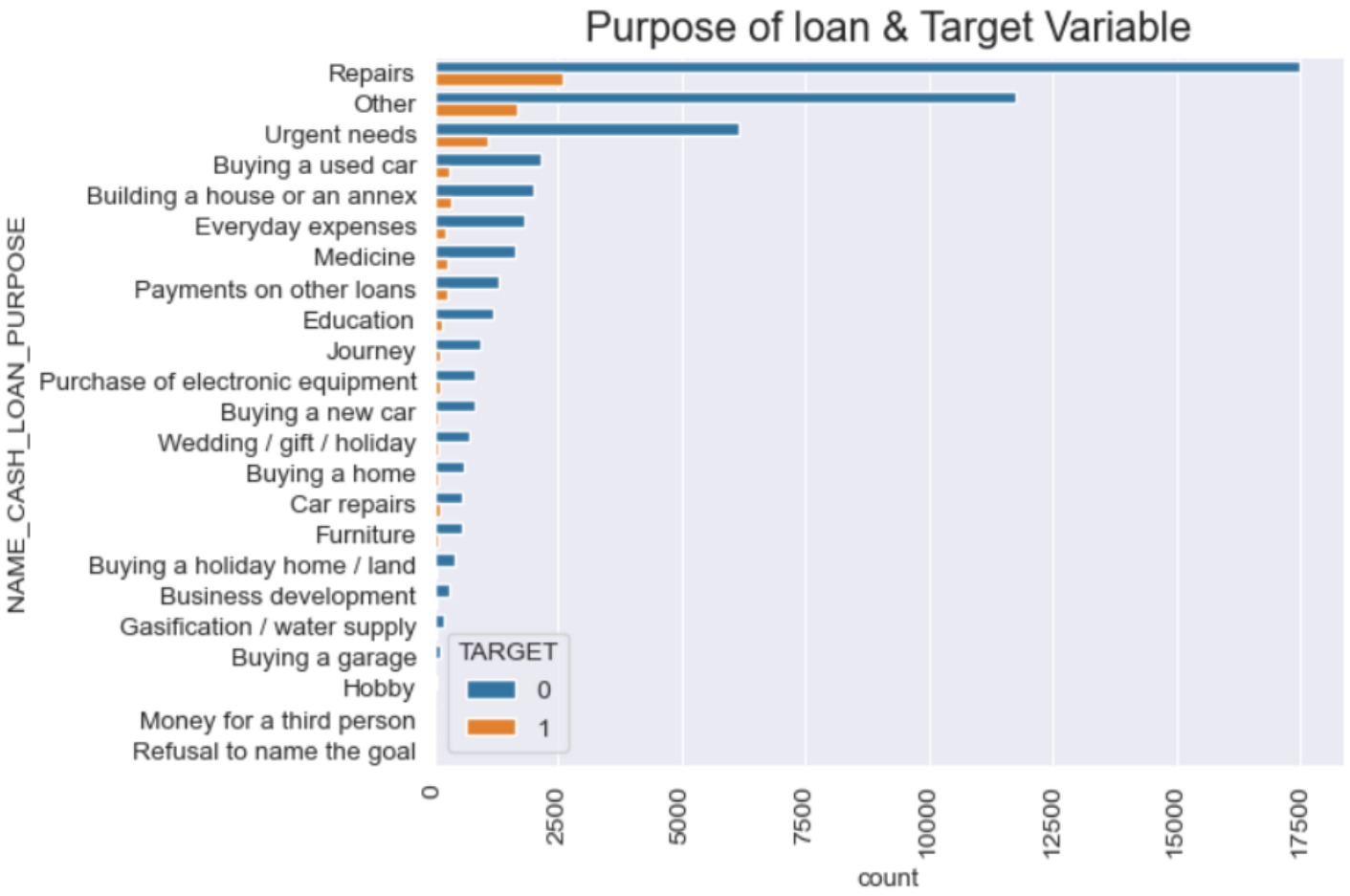
8. Merging the two datasets on SK_ID_CURR

9. Graphical Presentation on merging datasets

10. Observations derived from the graphs.

11. Conclusions and recommendations

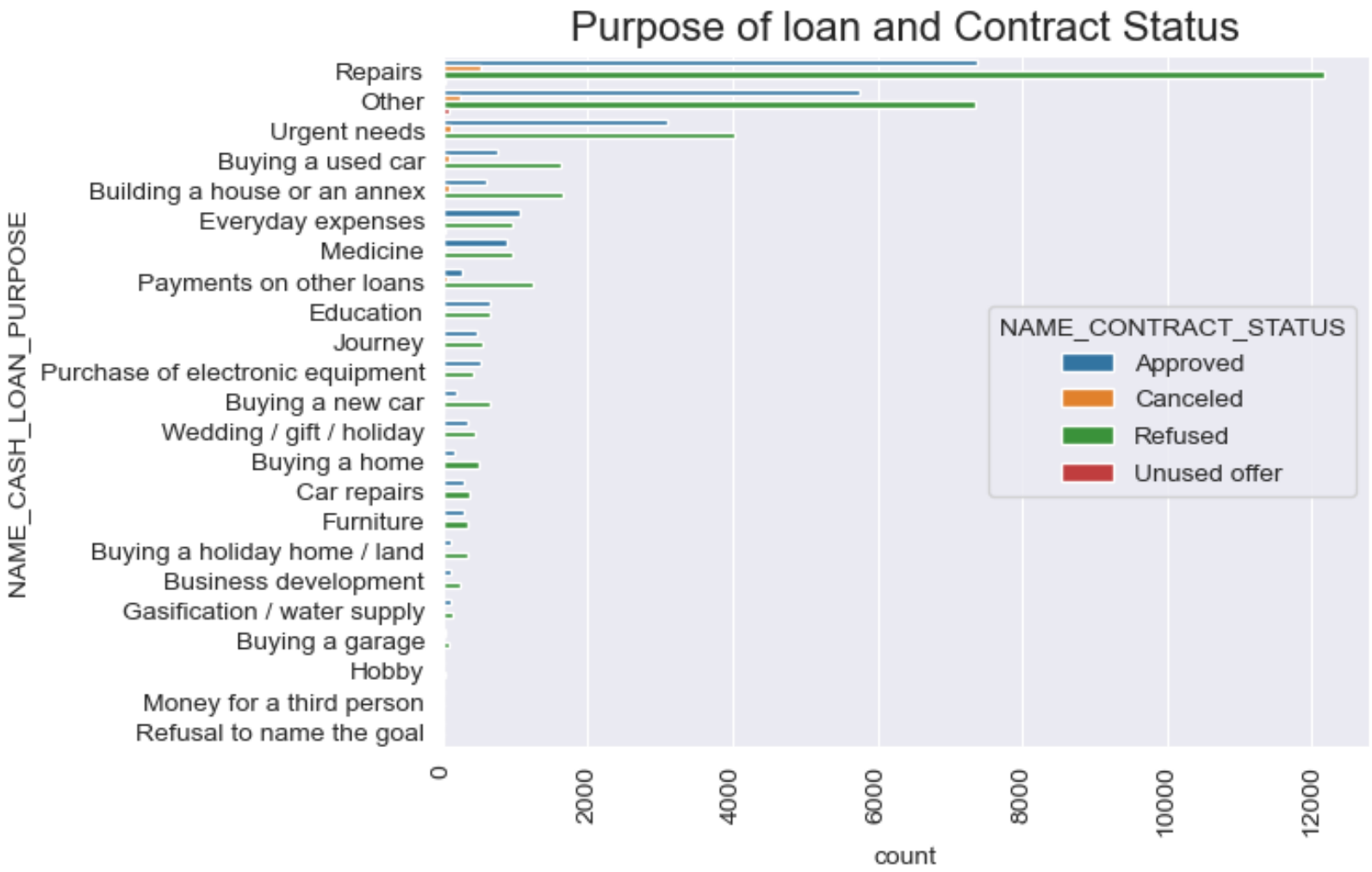
05. Purpose of Loan Vs Target Variable Graphical Representation



Observation from the above fig:

- Here we can see that the Repairs as a purpose of loan leads to more defaulters, hence we can avoid giving loans in this category.

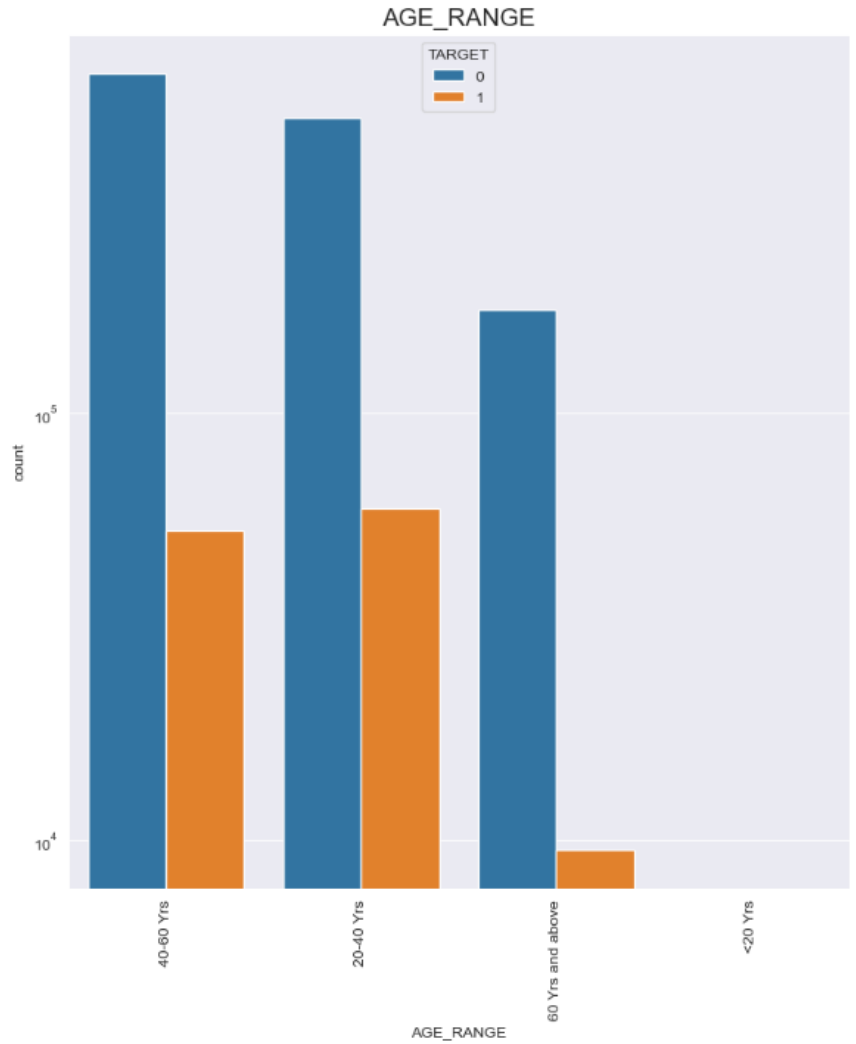
05. Purpose of loan Vs Contract Status

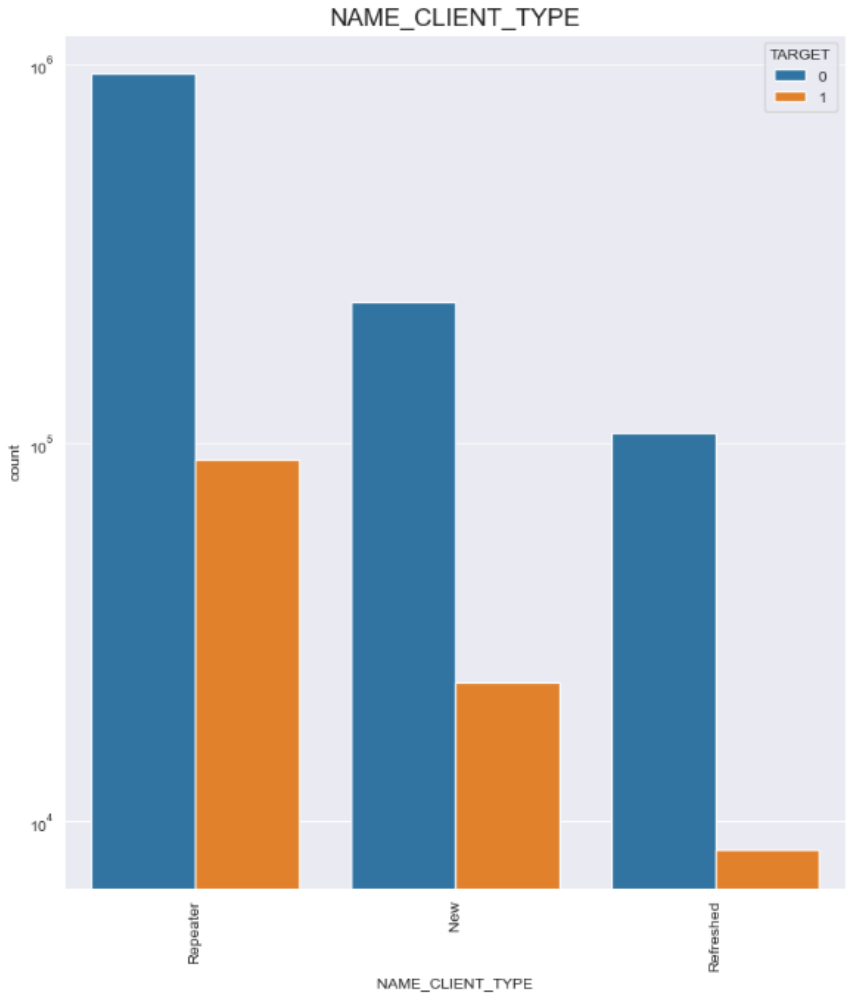


Observation from the above fig:

- Here we can see that the Approval for the Repairs as a purpose of loan is high and it will lead to defaulters

07. Bivariate analysis with TARGET, AGE and Client Type

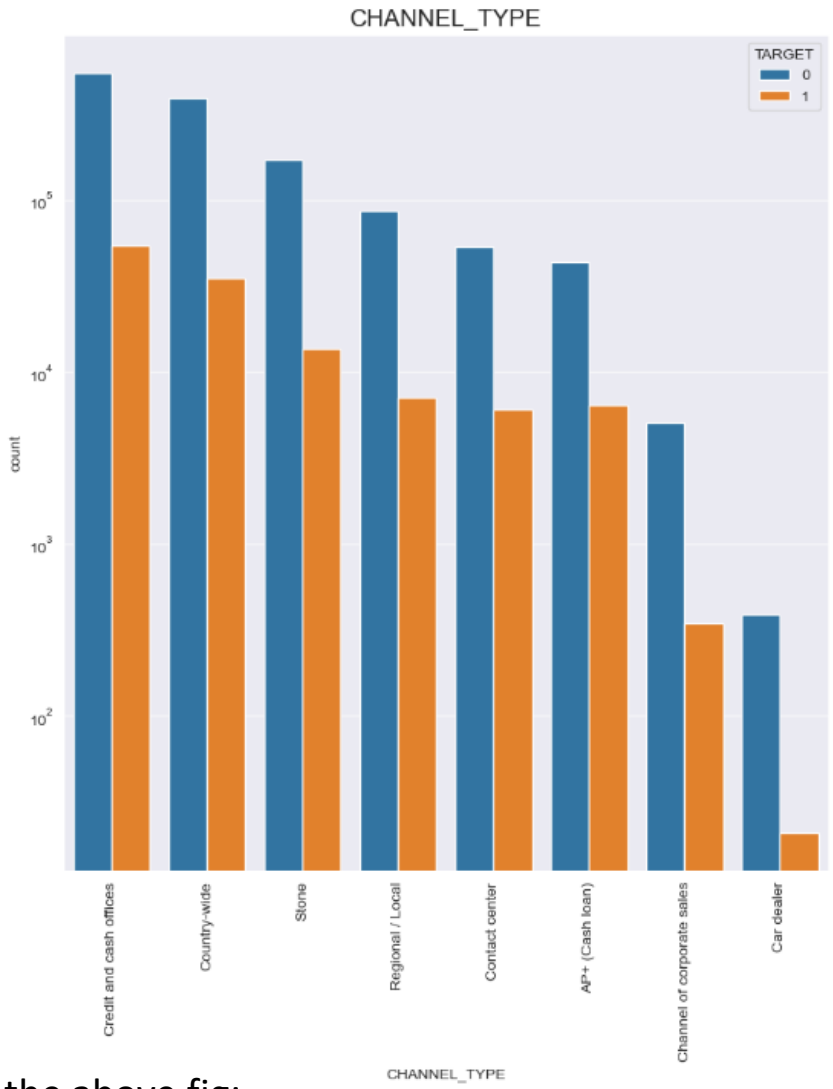




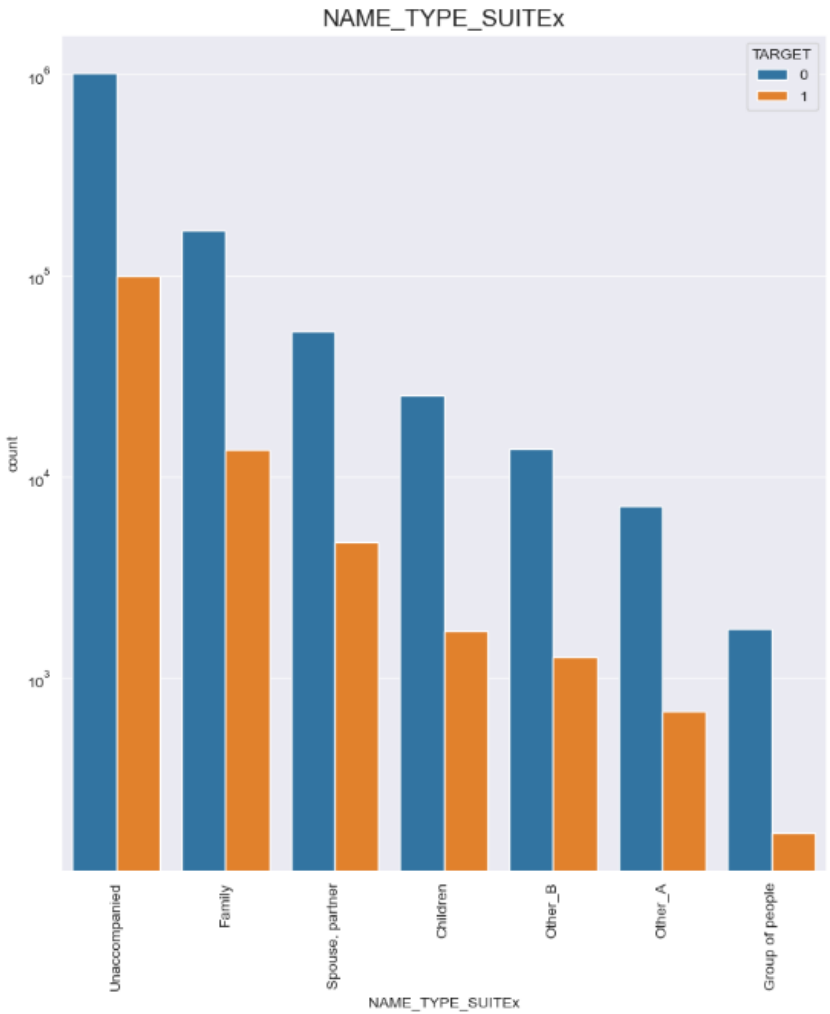
Observation from the above fig:
-There are more defaulters than non-defaulters in 20-40 years age range than 40-60 years range, also there is considerably less risk with clients of age 60 years and above.

Observation from the above fig:
- Repeaters have more defaulters than New applicants, while Refreshed applicants have the least defaulters.

07. Bivariate analysis with TARGET, Channel Type and Name Suit Type

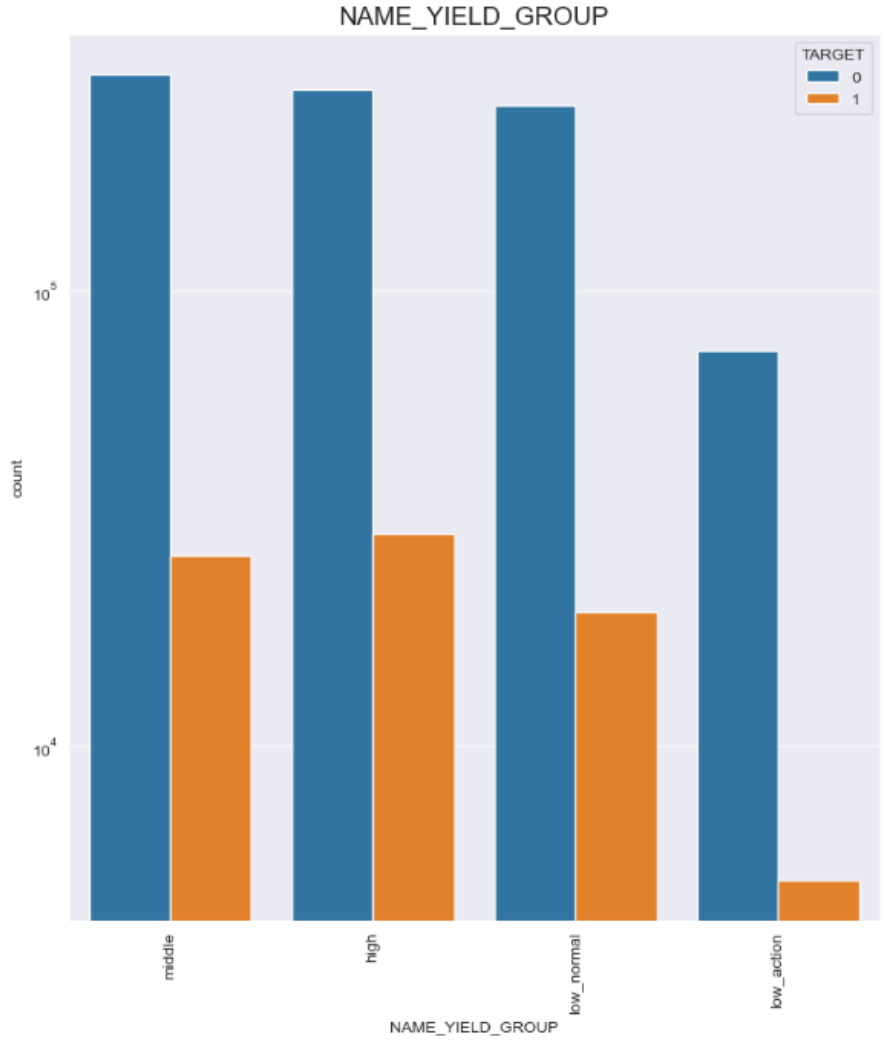


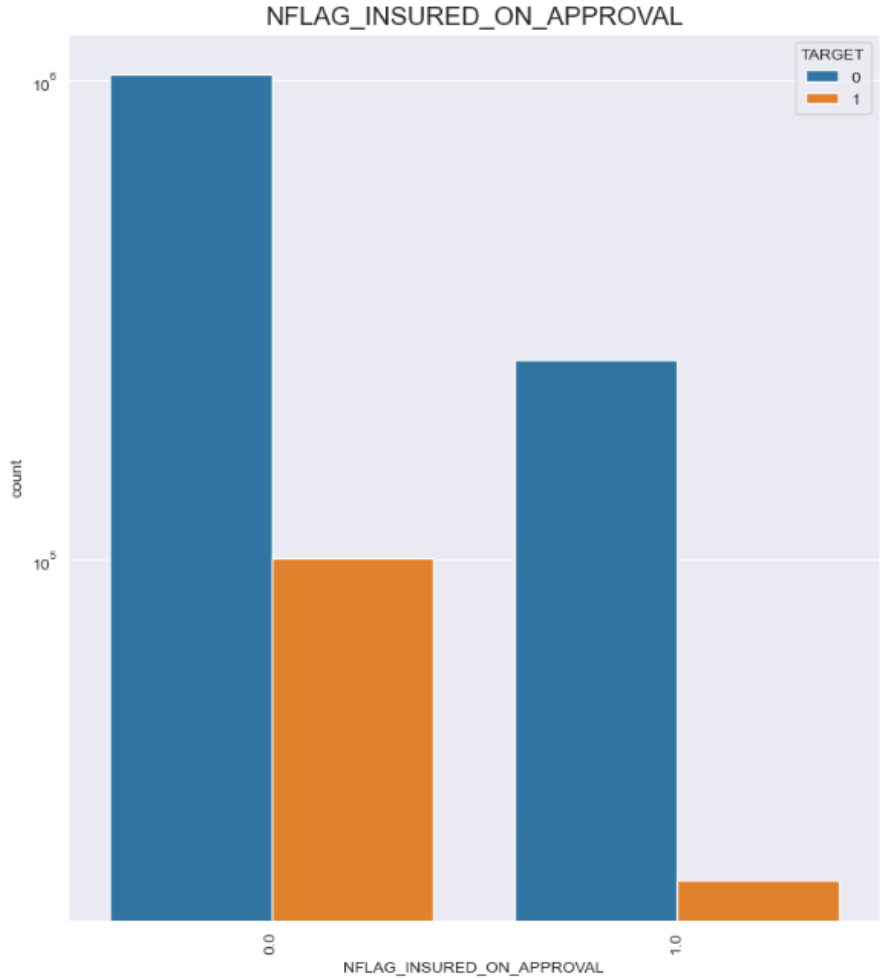
Observation from the above fig:
- We are getting more clients from Credit and cash offices at the same time defaulters are also high.



Observation from the above fig:
- Those applicants who are accompanied with family, spouse, children are less risky.

07. Bivariate analysis with TARGET, Name Yield group and NFLAG Insured Approval

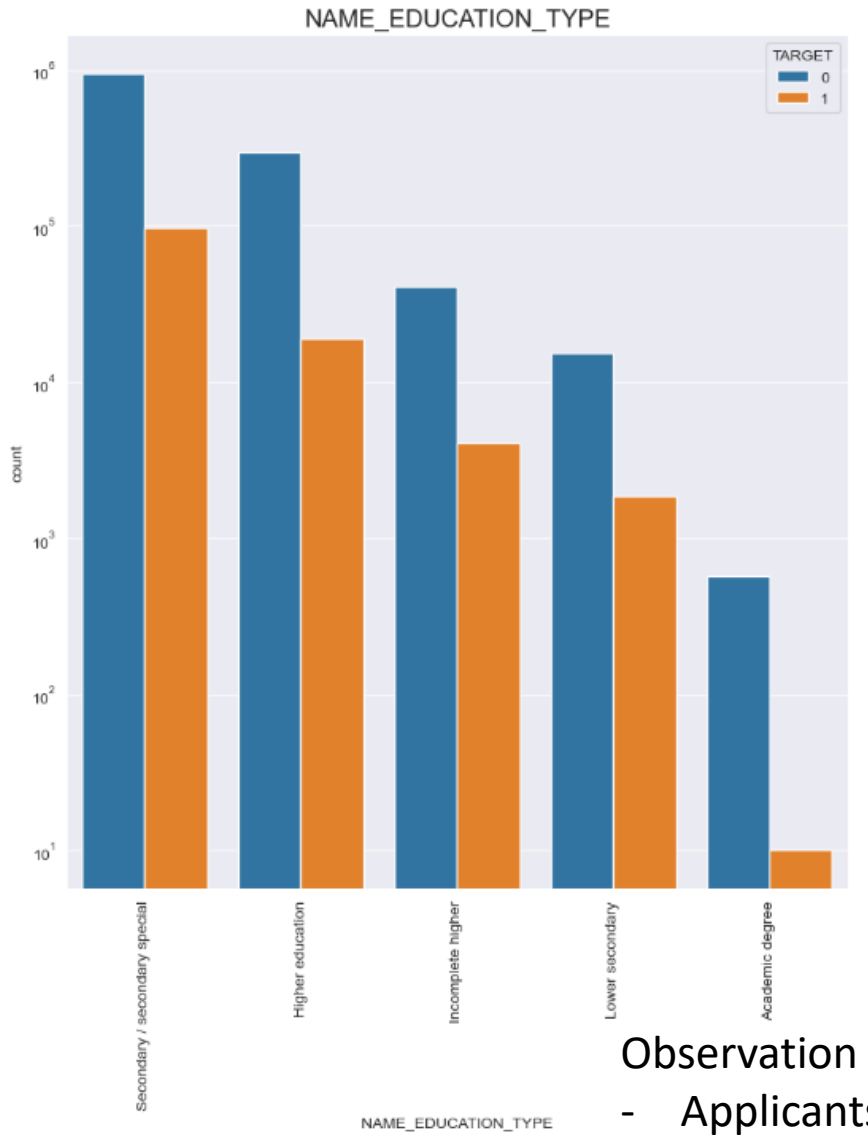


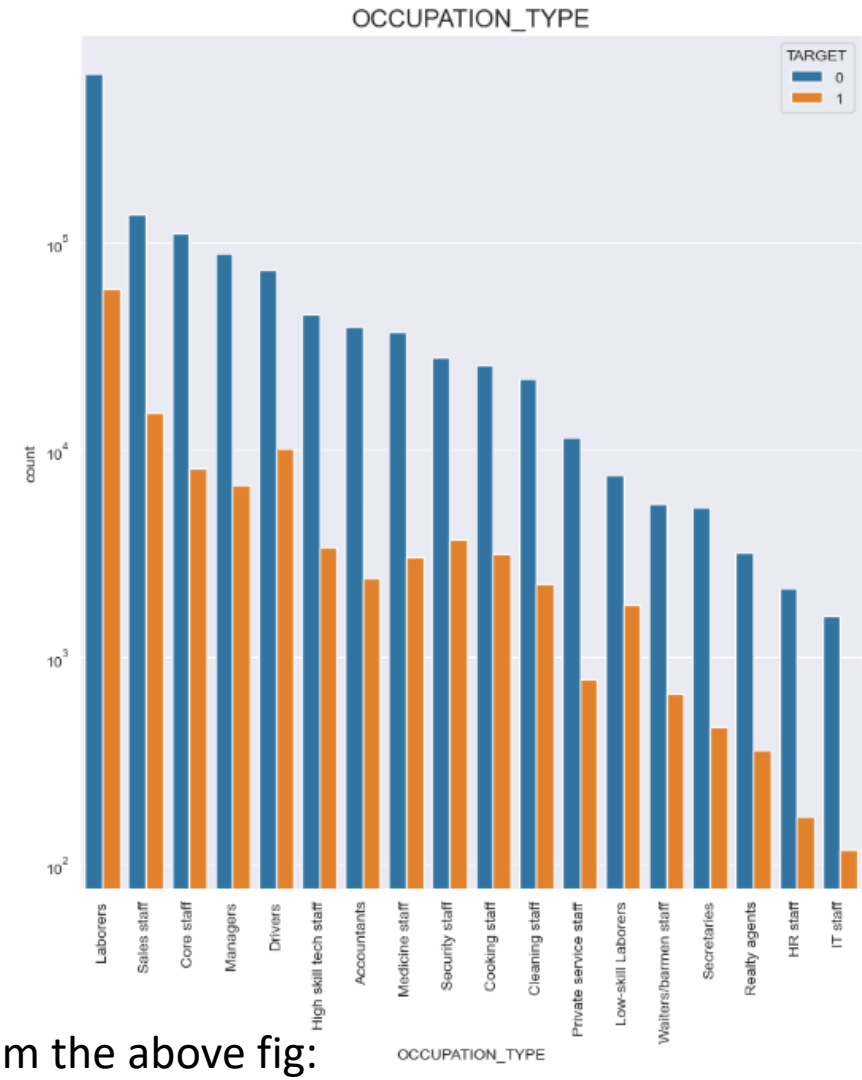


Observation from the above fig:
- High interest rates lead to slightly more defaulters while low_action interest rates are the least risky.

Observation from the above fig:
- Insured applicants do not tend to be defaulters more often.

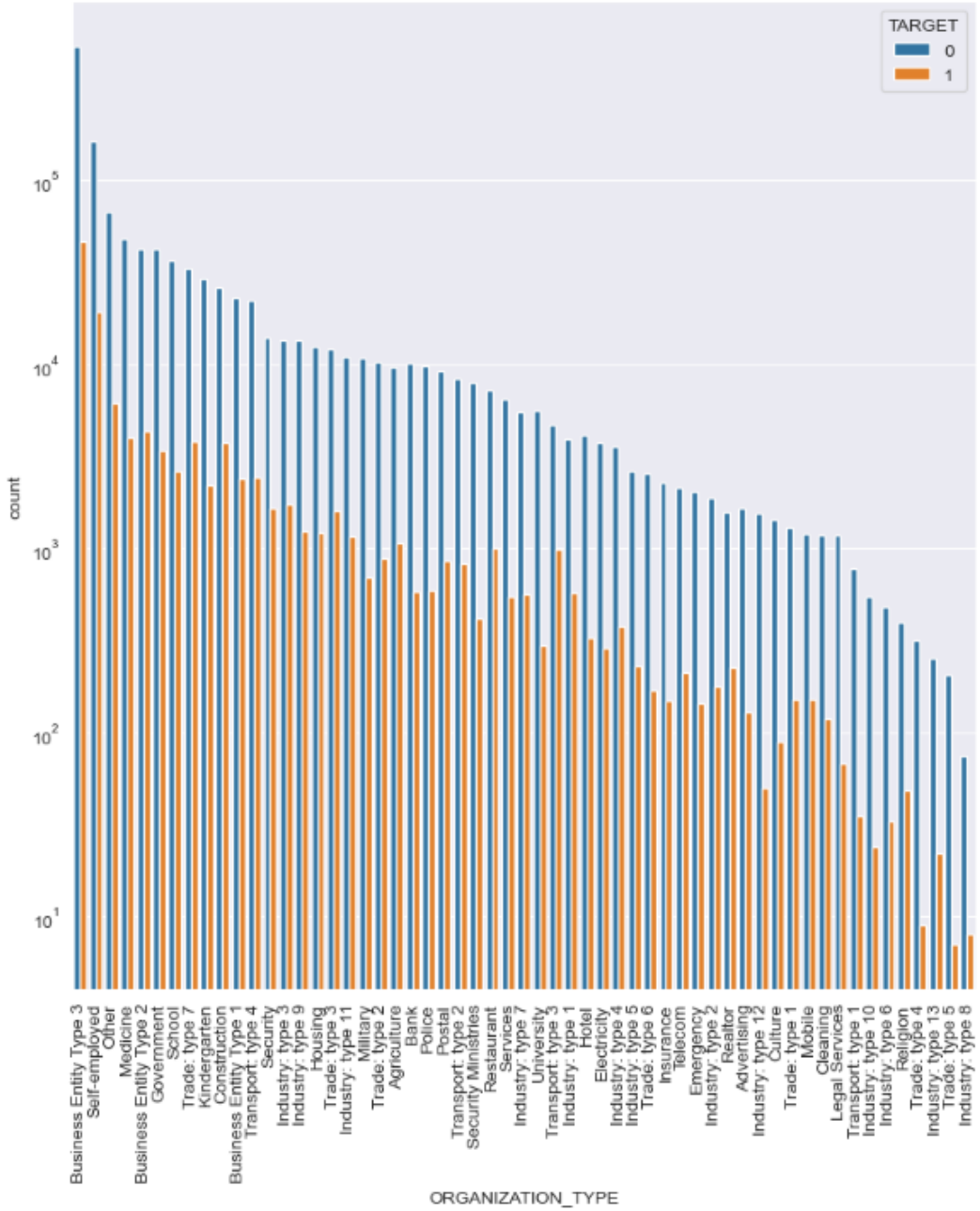
07. Bivariate analysis with TARGET, Education Type and Occupation Type





Observation from the above fig:

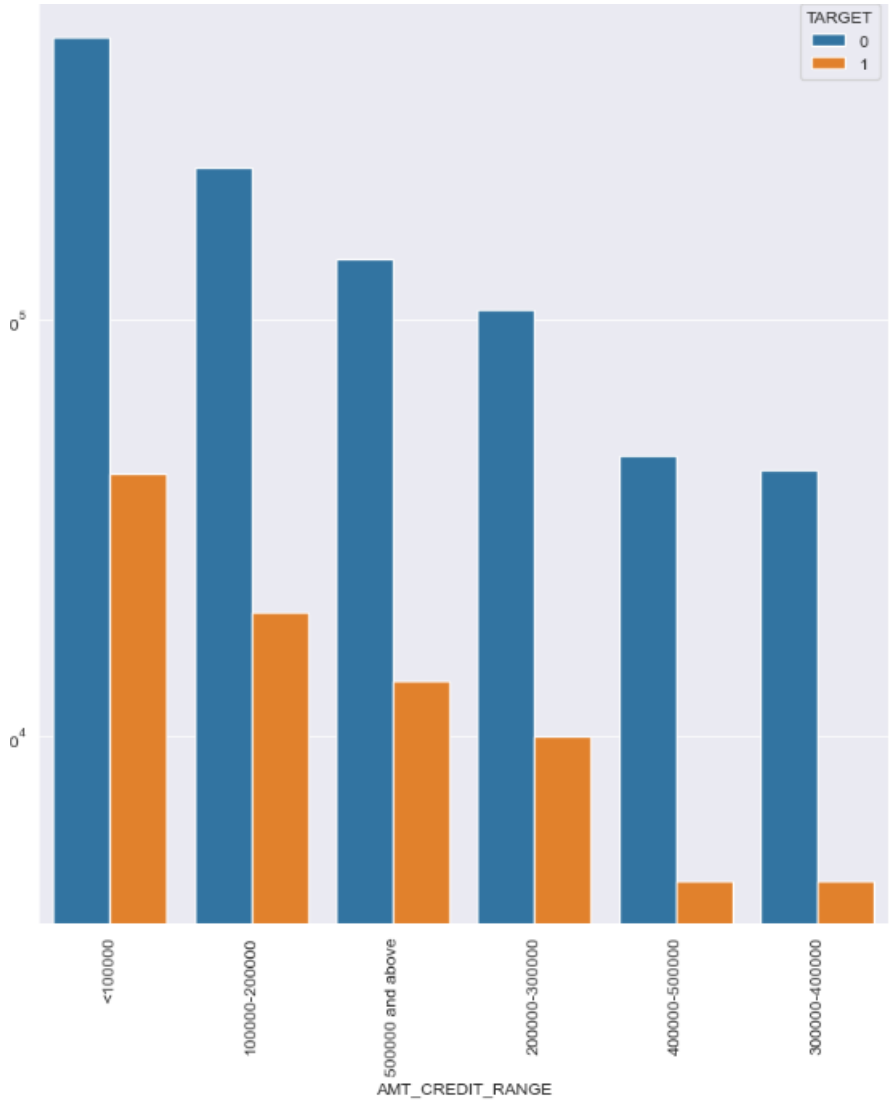
- Applicants with Secondary education lead to more defaulters.
- Labourer as the occupation type has more defaulters.



07. Bivariate analysis with TARGET and Organization Type

Observation from the above fig:

- Business Entity Type 3 and Self employed as organization type lead to more defaulters.



07. Bivariate analysis with TARGET and AMT Credit Range

Observation from the above fig:
- Credit amounts <100000 leads to more defaulters i.e high credit is less risky.

08. Conclusions and recommendations

1. - Applicants with income type 'Working' have high no. of defaulters, so banks should be cautious while 'Business' and 'Student' have the least no. of defaulters.

2. - People having income range 100000-200000 have high number of loans and also have more defaulters than others, while income segment > 500000 has less defaulters.

3. - 'Females' apply more for loans and tend to be become defaulters slightly more often than males.

4. - People apply more for 'Cash Loan' than 'Revolving Loan', also defaulters ae higher in 'Cash Loan'

5. - There are more defaulters than non-defaulters in 20-40 years age range than 40-60 years range, also there is considerably less risk with clients of age 60 years and above.

08. Conclusions and recommendations

6. Repeaters have more defaulters than New applicants, while Refreshed applicants have the least defaulters.

7. High interest rates lead to slightly more defaulters while low_action interest rates are the least risky.

8. Insured applicants do not tend to be defaulters more often.

9. Labourers as the occupation type has more defaulters

10. Applicants with Secondary education lead to more defaulters.

11. Credit amounts <100000 leads to more defaulters i.e high credit is less risky.

Thank You